DIAMOND PRICE PREDICTION

Uma Krishnan
CS DEPARTMENT
PES UNIVERSITY
BANGALORE, INDIA
umakrishnan2002@gmail.com

Srushti V Redyy
CS DEPARTMENT
PES UNIVERSITY
BANGALORE, INDIA
srushtivreddy@gmail.com

Tanuj Sadasivam
CS DEPARTMENT
PES UNIVERSITY
BANGALORE,INDIA
tanujsadasivam@gmail.com

***Abstract*—The Diamond market serves as a brilliant investment as loose diamonds are brilliant assets due to their reliability. They are reliable as their market is only modestly volatile. Thus, they are very minimally subjective to depreciation and are marginally vulnerable to inflation which generally returns good profits. Using data about diamonds and their features, we intend to analyse the price of Diamonds in the current day on the basis of Carat, Cut, Clarity, Colour, Depth, Table, their dimensions and how these features affect their price in the market.**

Keywords—volatile, dimensions, price, reliability, investment.

I. INTRODUCTION

There are numerous models and applications currently in the market which are used for predicting the future price of these metals .
Machine learning algorithm are generally classified into two main categories, namely supervised and unsupervised machine learning methods.
Some of the popular machine learning models which have been used for forecasting the future prices of Gold and Diamond are regression models like linear regressor, Random Forest, and ensemble techniques .
First, PCA, recursive feature elimination, and Chi-square test have been used to eliminate the correlation among features and obtain the best subset of features • Second, ensemble models based the strength of random forest and linear regression are used for predicting the future prices of Gold and Diamond The remainder of this paper is arranged as follows.

II. APPROACH USED

Ensemble learning is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.

Although there are a seemingly unlimited number of ensembles that you can develop for your predictive modeling problem, there are three methods that dominate the field of ensemble learning. So much so, that rather than algorithms per se, each is a field of study that has spawned many more specialized methods.

The three main classes of ensemble learning methods are bagging, stacking, and boosting, and it is important to both have a detailed understanding of each method and to consider them on your predictive modeling project.

.

III. SUMMARY OF RESULTS REPORTED

As a result of training various regression models and analyzing the results of the algorithms such as Linear Regression, Support Vector Regression, Random Forest Regression Decision Tree, Huber, Passive Aggressive, Bayesian, Extra Tree, K-neighbor, XGBoost and CatBoost, it can be concluded that CatBoost Regression is the most suitable algorithm for diamond price prediction having the highest R2 score of 0.9872 with comparatively lower RMSE and MAE.

IV. LIMITATIONS REPORTED

Noisy data and overfitting data were reported in 2 papers. Apart from this, there were no specific limitations reported. There was also no Lacuna in the evaluation that we inferred.

V. PRPOSED PROBLEM STATEMENT

Our aim is to implement a machine learning model to proficiently predict diamond prices based on attributes such as Carat, Cut, Clarity, Colour, Depth, Table, their dimensions in the present day to the highest degree of accuracy and certainty with the least conceivable error. We use Exploratory Data Analysis and PCA, as well as Ensemble Learning methods to provide us with statistical information that will guide us as to how the prices are dependent on the formerly stated attributes as well as whether they may be vulnerable to fluctuation in price.

.

.

REFERENCES

José[1] data mining techniques like M5P, linear regression, and neural networks. M5P data mining model shows that this model possesses a high capacity to predict the diamond price. Using Dimensionality Reduction by high correlation proves that this M5P model is a useful technique to apply on a diamond dataset. e. Ling et al. [2] have used different ensemble models by explaining the types of classifiers contained in it and used the model for prediction. Ensemble models use the notion of voting, bagging and stacking techniques to enhance the prediction accuracy. In [3], the authors trained other models including Decision Tree Regression with accuracy 100.00, Random Forest Regression with accuracy 99.50, Linear Regression with accuracy 91.87, Gradient Boosting Regression with accuracy 90.38, Lasso Regression with accuracy 90.17, AdaBoost Regression the accuracy with 85.10, Elastic Net Regression with accuracy 81.22 and Ridge Regression with accuracy 80.12. The model that gives the highest accuracy is Decision Tree Regression
In [4], the authors applied several models and reports their performance in term of R2-score, a statistical metric of how close the data are to the fitted model and the higher the R2-score, the better the model fits your data. Random Forest

Regression achieved score of 0.982066, K-Neighbours Regression 0.959033, Gradient Boosting Regression 0.905833, AdaBoost Regression 0.882499, Linear Regression 0.881432, Lasso Regression 0.865866 and Ridge Regression 0.753705. Random forest algorithm gives the highest R2-score "98%".

VI. REFERENCES

[1] José M., "Implementing data mining methods to predict Diamond prices" Peña Marmolejos Graduate School of Arts and Sciences, Fordham University, Int'l Conf. Data Science-ICDATA'18. https://csce.ucmss.com/cr/books/2018/LFS/CSREA2018/ ICD8070.pdf

[2] X. Ling, W. Deng, C. Gu, H. Zhou, C. Li, and F. Sun, "Model ensemble for click prediction in bing search ads," in Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017, pp. 689–698

[3] Diamond Price Modelling. (n.d.). Retrieved 5 December 2019, from https://kaggle.com/tobby1177/diamond-pricemodelling

[4] Diamonds In-Depth Analysis. (n.d.). Retrieved 5 December 2019, from https://kaggle.com/fuzzywizard/diamonds-indepth-analysis